**\section{Introduction}**

All research studies require the researchers to conduct experiments so the study results provide a basis confidently supporting or refuting a particular hypothesis. A carefully considered experimental plan is essential, and this important discipline of study is called \emph{experimental design}. The initial foundations of the statistical approach to experimental design were laid out by \cite{Fisher1935} in the field of agriculture, but the approach is now applicable to almost all sciences. \citeauthor{Fisher1935}'s main focuses were on the principles of comparison, randomisation, replication, blocking, orthogonality and the use of factorial treatments in connection with experimental design. Researchers wishing to obtain results that will stand up to scrutiny must implement these concepts in carefully designed experiments.

Carefully thought out experimental design offers several advantages. First, experiments are used to provide scientific evidence, so researchers should design and conduct their experiment so as to ensure reasonable confidence that it will yield accurate conclusions, i.e.\

the experimental conclusion must be \emph{valid} \citep{Maxwell2004}. Second, good experimental design should increase the information obtained per experiment versus an ad hoc approach, because such experiments can protect the inability to distinguish effects of interest from nuisance sources of variation, i.e.\ \emph{confounding}. Furthermore, the experimental design should provide an organised approach to conduct experiments, analyse datasets and interpret the results. Good experimental design can also ensure the experimental findings are reproducible, which facilitates communication between statisticians and researchers \citep{Doyle2009}.

This thesis focuses on the \emph{quantitative high-throughput biotechnologies experiment}. This experiment involves measuring changes in the abundance of certain intracellular molecular species, such as genes, proteins or metabolites, and searching for links between these changes and the presence or severity of conditions of interest. However, in-vivo measurement is generally impossible without high-throughput biotechnologies that can simultaneously be tested on a sample that comprises numerous candidate molecular species \citep{Janzen2002}. One example is the \emph{proteomics experiment} which involves the identification and quantification of the entire complement of proteins at a given condition within a cell. Protein abundances are measured using the Multi-dimensional Protein Identification Technology (MudPIT).

Proteomics experiments generally have a \emph{two-phase} structure. The organisms are first perturbed by the experimental conditions of interest, i.e.\ \emph{Phase 1 experiment}. Since protein abundance cannot be measured directly from the organisms, the organ must be harvested and the proteins extracted for the subsequent laboratory based experiment, i.e. the \ \emph{Phase 2 experiment}, which uses mass spectrometry to measure abundance. Thus, such experiments are also known as \emph{two-phase experiments}.

Due to the high variation between different MudPIT experiments, a \emph{multiplexing} technology is introduced that allows the simultaneous analysis of multiple samples, for example isobaric Tags for Relative and Absolute Quantitation (iTRAQ$^{\rm TM}$) \citep{Ross2004}. Another advantage of multiplexing is that it reduces overall experimental costs. However, complications in the experimental design arise in the design of the Phase 2 experiment, specifically in how the samples should be measured using this multiplexing technology.

This chapter establishes some general insights into the two-phase experiments. Section~\ref{sec:introTwoPhase} to \ref{sec:brien2011} then describes the evolution of the methods surrounding the two-phase experiments over recent decades. Moreover, Section~\ref{sec:proteomicExpt} details the biological background of the MudPIT-iTRAQ$^{\rm TM}$ experiments. Finally, Section~\ref{sec:overview} presents a brief overview of this thesis.

**\section{The introduction of two-phase experiments by McIntyre}**

**\label{sec:introTwoPhase}**

Two-phase experiments were first introduced by \cite{McIntyre1955}, who investigated the effects of four light treatments on the synthesis of tobacco mosaic virus in tobacco leaves. The Phase 1 experiment comprised two $4 \times 4$ arrays consisting of four leaves taken from each of eight infected plants. The four light treatments were then assigned to the plants and leaves such that each treatment occurs only once within each row and column in each of two $4 \times 4$ square arrays. This assignment is also known as \emph{Latin square design} \citep{Bailey2008}. The Phase 1 experiment thus yielded 32 observations. However, the virus content of each leaf could not be measured directly from the test plants used in the Phase 1 experiment. Therefore, in the Phase 2 experiment the virus content of each leaf was assayed by expressing sap and inoculated using the half leaves of the assay plant, where the countable lessons are appeared.

The Phase 2 experiment comprised four $4 \times 4$ arrays consisting of four leaves taken from each of 16 assay plants. Since each leaf was further divided into two halves, the Phase 2 design can take the form of two sets of Latin square designs superimposed on each other, an arrangement also known as the \emph{Greaco-Latin square design}. Hence, a total of 128 samples are measured in the Phase 2 experiment. Given that the Phase 1 experiment involved 32 samples, each sample was replicated four times in the Phase 2 experiment. This experiment provides a good example of a two-phase experiment that requires two different experimental designs, where the first design prepares a set of test plants that were perturbed by the light treatments of interest while the second design allows the measurements of the lesions on the assay plants that correspond to the virus content of a certain treatment and test plant from the Phase 1 experiment.

\cite{McIntyre1955} presented two important principles in the design of the two-phase experiment. Although the measurement cannot be made based on the Phase 1 experiment, \cite{McIntyre1955} stated that the replication in the Phase 1 experiment is essential. This is because the theoretical variances of treatment differences must be obtainable to measure the consistency of performance and to theoretically obtain a valid test for the significance. This replication is also known as \emph{biological replication}, i.e.\ samples are taken from each of animal in the healthy and diseased groups.

If multiple measurements on each sample from the Phase 1 experiment are not made in the Phase 2 experiment, the variance of treatment differences will comprise both biological and technical variation, i.e.\ the biological variation will be confounded with the measurement error. This multiple measurement is also known as \emph{technical replication}, i.e.\ measuring the subsamples from each Phase 1 biological sample. \cite{McIntyre1955} stated that this replication is only required if the Phase 2 experiment introduces large and uncontrollable variation. For the light treatment example mentioned above, each treatment is replicated four times in Phase 1. Additionally, each sample in the Phase 1 experiment is measured four times in the Phase 2 experiment. Thus, both types of replication are used in the light treatment experiment.

The construction of the ANOVA table with EMS is shown to usefully illustrate the linear combination of the variance components because it allows the estimation of the error variance for the treatment effects. \cite{McIntyre1955} discussed three further observations using these error variances and based on the light treatment experiment. First, he demonstrated that if the Phase 1 design remains unchanged while the Phase 2 design is modified, the error variances for the treatment comparisons change. This is because the different designs cause different combinations of the variance components. Second, if the design is modified to have more technical replicates, the error variance for the treatment effects can be reduced, but the time required to complete the experiment is increased. Third, if the design is modified to eliminate biological replication, the error variance for the treatment effects increases. This is because the error variance includes the variation among leaves from single plants from both the Phase 1 and 2 experiments.

**\section{New ANOVA by \cite{Curnow1959}}**

\cite{Curnow1959} revisited \citeauthor{McIntyre1955}'s light treatment experiment and designed a new ANOVA table. The new ANOVA table is better than that presented by \cite{McIntyre1955}, because it shows the decomposition of the total sum of squares (SS) into each source of variation associated with each experimental treatment/block factor. Table~\ref{tab:Curnow1959}, which includes only the treatment and residual SS, shows that the treatment effects can be estimated both Between and Within Leaves in Assay Plants. The components in EMS are $\sigma\_{\epsilon}^2$ which denotes the variation between halves within leaves of assay plants, $\sigma\_{\Delta}$ which denotes the variation between leaves in assay plants, $\sigma\_{L}$ which denotes the variation between leaves in test plants and $\theta$ which is the differences between the treatments. The first estimate of the treatment effects contains $\sigma\_{\Delta}^2$, whereas the second estimate does not. \cite{Curnow1959} labels these first and second estimations of the treatment effects \emph{sums} and \emph{differences} analyses, respectively, which are equivalent to \emph{inter} and \emph{intra}-block analyses. \cite{Curnow1959} then showed how to combine the inter- and intra-block analyses using the weights computed from the variance of treatment estimates.

\begin{table}[ht]

\centering

\caption{Partial theoretical ANOVA table by \cite{Curnow1959}.}

\begin{tabular}{lrl}

\toprule

\multicolumn{1}{l}{\textbf{Source of Variation}} & \multicolumn{1}{l}{\textbf{DF}} & \multicolumn{1}{l}{\textbf{EMS}}\\

\midrule

Between Leaves in Assay Plants & & \\

\quad Treatment & $3$ & $\sigma\_{\epsilon}^2 + 2\sigma\_{\Delta}^2 + 2\sigma\_{L}^2 + 16\theta$ \\

\quad Residual & $15$ & $\sigma\_{\epsilon}^2 + 2\sigma\_{\Delta}^2 + 2\sigma\_{L}^2$ \\ \hline

Within Leaves in Assay Plants & & \\

\quad Treatment & $3$ & $\sigma\_{\epsilon}^2 + 2\sigma\_{L}^2 + 16\theta$ \\

\quad Residual & $15$ & $\sigma\_{\epsilon}^2 + 2\sigma\_{L}^2 $ \\

\bottomrule

\end{tabular}

**\label{tab:Curnow1959}**

\end{table}

In contrast with \cite{McIntyre1955}, who performed the unweighed pooled estimation, the estimations of treatment in the Between and Within Leaves in Assay Plants are combined in the ANOVA table in Table~\ref{tab:McIntyre1955}.

The total degrees of freedom (DF) of the light treatment experiment is 127, but the total DF from the original ANOVA table by \cite{McIntyre1955} is 155, which is incorrect. Additionally, \cite{McIntyre1955} also omitted the 3 DF for the Greaco-Latin Squares in the original ANOVA table. Therefore, the new ANOVA table by \cite{Curnow1959} took account of all 127 DF from the light treatment experiment, and thus provides a first step towards developing a better analytical procedure for the two-phase experiments.

\begin{table}[ht]

\centering

\caption{Partial theoretical ANOVA table by \cite{McIntyre1955}.}

\begin{tabular}{lrl}

\toprule

\multicolumn{1}{l}{\textbf{Source of Variation}} & \multicolumn{1}{l}{\textbf{DF}} & \multicolumn{1}{l}{\textbf{EMS}}\\

\midrule

Treatment & $3$ & $\sigma\_{\epsilon}^2 + \sigma\_{\Delta}^2 + 4\sigma\_{L}^2 + 32\theta$ \\

Residual & $15$ & $\sigma\_{\epsilon}^2 + \sigma\_{\Delta}^2 + 4\sigma\_{L}^2$ \\

\bottomrule

\end{tabular}

**\label{tab:McIntyre1955}**

\end{table}

**\section{Non-orthogonal block structure in the two-phase experiment}**

\cite{Wood1988} discussed the non-orthogonal block structure of the two-phase experiments, which occurs when the block factor from the Phase 1 experiment is confounded with that from the Phase 2 experiment. The orthogonality can be measured by the \emph{efficiency factor} which is the proportion of treatment information presented for the intra-block analysis \citep{Yates1936}. The efficiency factor thus should also be used to measure the block information for cases involving non-orthogonal block structure as shown by \citep{Wood1988}.

This thesis studies both the non-orthogonal block and treatment structures, which can occur in two different ways. The first is that the DF of block effects from the Phase 1 experiment are split into inter- and intra-blocks. This can be shown in the ANOVA table given by \cite{Curnow1959} where the 6 DF associated with the leaf positions of the test plants are split into two SS, both estimated with 3 DF. The first SS is confounded with the whole leaves of the assay plants while the second is orthogonal to them. The efficiency factor for both SS can be shown to be 1. The second origin is that the DF of block effects from the Phase 1 experiment remain intact. This can be shown in the artificial example by \cite{Wood1988} presented in Table~\ref{tab:Wood1988}. In the ANOVA, all three DF associated with the units SS of the Phase 1 experiment are estimated in both the Between and Within Blocks. The separation for this case is then measured using the efficiency factor. Both types of confounding should be minimised, because they can affect the conduct of the test for the treatment effects. Chapters 3 and 4 describe the methods used to find the optimal design and also remain aware of the issues in the non-orthogonal block structures and how they affect the ANOVA.

\begin{table}[ht]

\centering

\caption{Phase 2 design showing the allocation of Phase 1 units to Phase 2 blocks in the artificial example by \cite{Wood1988}.}

\begin{tabular}{ccccc}

\bf Block & 1 & 2 & 3 & 4 \\ \hline

& 1 & 2 & 3 & 4 \\

& 1 & 2 & 3 & 4 \\

& 2 & 3 & 4 & 1 \\ \hline

\end{tabular}

**\label{tab:Wood1988}**

\end{table}

**\section{Constructing the analysis of variance table}**

**\label{sec:Brien1983}**

Neither \cite{McIntyre1955} nor \cite{Curnow1959} described how their ANOVA tables were determined. \cite{Brien1983} thus presented a generalised procedure for determining the ANOVA table for both single-phase and two-phase experiments. Notably, the presented ANOVA consists only of the decomposition of the total DF into each source of variation, without the EMS. This section briefly describes the procedure presented by \cite{Brien1983} and simultaneously introduces some basic terminologies used in experimental design.

Before constructing the ANOVA table, the overall structure of the experiment must be determined. The first step is to identify the treatment and block factors in the experiment and the observational unit. The observational unit is the smallest measured quantity of experimental material \cite{Bailey2008}. The second step is to divide the factors into different sets, which \cite{Bailey2008} called \emph{tiers}. The first tier consists of the factors that jointly identify the observational unit in the absence of randomisation. \emph{Randomisation} is an important concept in the experimental design that involves the assignment of one set of objects to another. The main purpose of randomisation is to enable researchers to obtain a data set with minimal systematic bias. \cite{Nelder1965A} also referred to these first tier factors as \emph{block} factors. The second tier factors are those level combinations that are directly associated with the factors at the first tier based on the randomisation. The second tier factors thus are what \cite{Nelder1965B} termed \emph{treatment} factors. The smallest unit in the first tier to which the treatment can be applied is also known as the \emph{experimental unit}. The last step is to determine the relationships between the factors within each tier. \cite{Wilkinson1973} developed a symbolic syntax to represent the block and treatment structures in an experiment. \cite{Brien1999} termed this representation a \emph{structure formula}. \citeauthor{Wilkinson1973}'s syntax was originally developed to generate and analyse the ANOVA models in the GenStat statistical analysis program, but is now widely used in many statistical packages. Two basic operations described by \cite{Wilkinson1973} are used to represent block and treatment structures, namely \emph{crossing} denoted by an asterisk, $\*$, and \emph{nesting}, denoted by a slash, $/$.

Two-phase experiments involve three tiers of factors, two of block factors and one of treatment factors. Consequently, two-phase experiments are also known as \emph{multi-tiered experiments}. Tiers 1 and 2 comprise block factors from the Phase 2 and 1 experiments, respectively. Tier 3 contains the treatment factors of the overall experiments.

The ANOVA table can be obtained once the structural formula for every tier of the experiment is determined. The first step is to expand the structure formulae for each tier, which yields a set of terms based on the rules described by \cite{Wilkinson1973}. The terms from the first tier form an initial structure of the ANOVA table. Meanwhile, the terms from the second tier are included in the table under the terms of the first tier with indentation, and only if the two terms from different tiers are confounded. The confounding between the terms can be examined using the contrasts generated from each term. The two-phase experiment includes an additional tier. Therefore, the terms from tier 3 are grouped with those from tiers 1 or 2, with another indentation in the event of confounding between terms from the pair of grouped tiers.

Using a wine-evaluation experiment described by \cite{Brien1983}, the factors of the field experiment are Blocks, Plots and Treatments with $b$, $p$ and $p$ levels, respectively. The factors of the tasting experiment are Tasters and Sittings with $t$ and $bp$ levels, respectively. The observational unit is the wine given to a taster at a particular sitting. The level combination of Taster and Sitting factors cannot be randomised; thus, Taster and Sitting factors formed the first tier. The level of Block and Plot factors combination is randomised to sittings within each taster; thus, the Block and Plot factors formed the second tier. Finally, the treatment is randomised to the plots within each block; thus, the treatment factor forms the third tier. The structure formulae of Phase 2 and 1 block and the treatment tiers can then be written as

\begin{eqnarray}

**\label{eq:stru1}**&&\mathrm{Taster/Sitting},\\

**\label{eq:stru2}**&&\mathrm{Block/Plot}

\end{eqnarray}

and

\begin{equation}\**label{eq:stru3}**

\mathrm{Treatment.}

\end{equation}

The structure formulae in (\ref{eq:stru1}) and (\ref{eq:stru2}) can then be expanded to

\begin{equation}\**label{eq:stru1expanded}**

\mathrm{Taster + Taster.Sitting}

\end{equation}

and

\begin{equation}\**label{eq:stru2expanded}**

\mathrm{Block + Block.Plot},

\end{equation}

respectively, where $\mathrm{Taster.Sitting}$ and $\mathrm{Block.Plot}$ denote Between Sittings Within Tasters and Between Plots Within Blocks, respectively. After studying the relationship between the terms from two different tiers, the ANOVA table, with only the decomposition of the DF, can be derived and is given in Table~\ref{tab:Brien1983}. The indention of the Treatment line indicates that the Treatment is confounded with $\mathrm{Block.Plot}$. For a more detailed description of these procedures see \cite{Brien1983}.

\begin{table}[ht]

\centering

\caption{ANOVA table of the wine-evaluation experiment described by \cite{Brien1983}.}

\begin{tabular}{lrl}

\toprule

\multicolumn{1}{l}{\textbf{Source of Variation}} & \multicolumn{1}{l}{\textbf{DF}} \\

\midrule

Between Tasters & $t-1$ \\

Between Sittings Within Tasters & $t(bp-1)$ \\

\quad Between Blocks & $b-1$ \\

\quad Between Plots Within Blocks & $b(p-1)$ \\

\quad \quad Treatment & $p-1$ \\

\quad \quad Residual & $(b-1)(p-1)$ \\

\quad Residual & $(bp-1)(t-1)$ \\

\bottomrule

\end{tabular}

**\label{tab:Brien1983}**

\end{table}

\cite{Brien1983} concluded by expressing that the randomisation between the factors of different tiers can alter the relationships among factors within each tier. Consider, for instance, an experiment arranged in a row-column design; the relationship between the Row and Column factors is \emph{crossed}. Consider the plots as the observational units, then suppose a randomised complete block design is to be superimposed on the plots, with treatment being randomised to the plots within each row. The treatment factor then becomes confounded with the column factor and the relationship between the row and column factors becomes \emph{nested}. Therefore, the experimental structure depends on both the innate physical structure of the experimental material and on how the randomisation is employed.

**\section{Complex two-phase experiments}\label{sec:Brien1999}**

A crucial aspect of the complex two-phase experiment is that at least one term in the structure formula of the third tier is non-orthogonal to the multiple terms in the structure formula of the second tier which themselves are non-orthogonal to terms in the structure formula of the first tier. Thus, there are two sets of efficiency factors, denoted by $E$, to describe the non-orthogonal relationship between the terms of different tiers, as mentioned by \cite{Wood1988}.

The procedures described in \cite{Brien1983} to construct the ANOVA table only provide the decomposition of the total DF. \cite{Brien1999} thus described an algorithm using two types of \emph{sweeping} operations to construct the ANOVA table which provides the estimation of the effects for each source of variation for any single- or two-phase experiment.

The field experiment described in Section~\ref{sec:Brien1983} is used to demonstrate the sweeping operations. The factor of this experiment consists of Blocks, Plots and Treatments with $b$, $p$ and $p$ levels, respectively. The expanded structure formulae of the block and treatment tiers are given in (\ref{eq:stru2expanded}) and (\ref{eq:stru3}), respectively. For this experiment to be complex, the Treatment term in (\ref{eq:stru3}) must be non-orthogonal to the Block term in (\ref{eq:stru2expanded}), which is an example of balanced incomplete block design.

The sweeping operations were introduced by \cite{Wilkinson1970} and \cite{Payne1977}. Every experiment generates $n$ responses which can be viewed as a data vector, commonly denoted by $\bm{y}$, with length $n$ in the $n$-dimensional space \citep{Payne1977}. The standard sweeping operation estimates the effects of a term from the expanded structure formula and then subtracts them from the current working vector which then becomes the working vector for the next sweep.

The field experiment described in Section~\ref{sec:Brien1983} is used to demonstrate the sweeping operations. The factors this experiment consists of are Blocks, Plots and Treatments with $b$, $p$ and $p$ levels, respectively. The expanded structure formulae of the block and treatment tiers are given in (\ref{eq:stru2expanded}) and (\ref{eq:stru3}), respectively. For the experiment to be complex, the $\mathrm{Treatment}$ term in (\ref{eq:stru3}) must be non-orthogonal to both the $\mathrm{Block}$ and $\mathrm{Block.Plot}$ terms in (\ref{eq:stru2expanded}), demonstrating balanced incomplete block design. Every experiment consists of a grand mean which automatically become part of the structure formulae \citep{Payne1977}. Thus, the first sweep is for the grand mean, i.e.\ removing the effects of grand mean, using the \emph{reanalysis sweep}, which can be expressed by

\begin{equation}

\mathbf{I} - E\_K \mathbf{P}\_K

\end{equation}

where $\mathbf{I}$ is the $n \times n$ identity matrix, while $E\_{G}$ and $\mP\_{G}$ are the efficiency factor and projection matrix associated with the grand mean. The $\mP\_{G}$ is given by

\[ {\K}({\K}'{\K})^{-1}{\K}'\]

where $\K$ denotes an $n \times n$ averaging matrix with all elements equal to ${n}^{-1}$. Generally, the reanalysis sweep is intended to sweep the non-orthogonal terms from the data vector; the sweep to remove the effect of the Block term from the data vector is given by

\[\mathbf{I} - E\_B \mathbf{P}\_B,\]

where $E\_{B}$ and $\mP\_{B}$ are the efficiency factor and projection matrix associated with the Block term. The second type of sweep operator is the \emph{pivotal sweep} which is used to estimate the effects of the associated terms. Thus, the sweep for estimating the effect of Treatment term is given by

\begin{equation}

E\_T \mathbf{P}\_T,

\end{equation}

where $E\_{T}$ and $\mP\_{T}$ are the efficiency factor and projection matrix associated with the Treatment term. This sweep is required for the terms of the structure formula for the first (block) tier before performing the same sweep for the terms in the structure formula for the second (treatment) tier that are confounded with it. Therefore, the sweeping sequence to estimate the effects of Treatment in Between Blocks for the field experiment is given by

\[E\_T \mP\_T E\_B \mP\_B (\I - E\_K \mP\_G)\bm{y},\]

which consists of two pivotal sweeps for the Treatment and Block terms, because the effects of Treatment are confounded with the Between Blocks, and a reanalysis sweep for the grand mean. The sweeping sequence for the estimation on the other effects of the experiment can also be derived from a combination of the reanalysis and pivotal sweeps. Thus, the ANOVA table can be constructed with the effects for each source of variation.

\cite{Brien1999} extended the sweep algorithm for the two-phase experiment by performing additional sweeps for terms from the structure formula of the third tier. Using the wine-evaluation two-phase experiment described in Section~\ref{sec:Brien1983}, the effects of Treatment in Between Plots Within Blocks and Between Sittings Within Tasters are given by

\[\mP\_T \mP\_P (\I - \mP\_B) \mP\_S (\I - \mP\_J) (\I - \mP\_G)\bm{y}, \]

where $mP\_P$, $mP\_S$ and $mP\_J$ denote the projection matrix associated with Plots, Sittings and Tasters, respectively. Since all block and treatment information is intact, all the efficiency factors are unity and are not written in the expression.

To summarize, \cite{Brien1999} demonstrated the use of the reanalysis and pivotal sweep operators to construct the ANOVA table. Chapter 2 further discusses these methods of constructing the ANOVA table.

**\section{Works by Brien and Bailey}**

\cite{Brien2006b} described how randomisation can be achieved for the multi-tiered experiment, which involves at least three tiers of factors. Thus, the multi-tiered experiment also covers the two-phase experiment. As \cite{Brien1983} pointed out, randomisation can affect the structure of the experiments. For an experiment that involves three tiers of factors, randomisation typically must be performed twice: (1) for the the allocation of treatments to experimental units in the Phase 1 experiment, and (2) for the allocation of experimental units from the Phase 1 experiment to the experimental units in the Phase 2 experiment. The randomisation procedure for the multi-tiered experiment thus is termed \emph{multiple randomisation} \citep{Brien2006b}.

\cite{Brien2006b} then compared and contrasted six types of multiple randomisation with three tiers of experiments, namely composed, coincident, independent, double, randomised-inclusive and non-randomised-inclusive multiple randomisations. These six multiple randomisations differ in both direction and the interaction of factors in different tiers. For a more detailed description of these multiple randomisations see \cite{Brien2006b}.

\cite{Brien2009, Brien2010} also published two papers on the aspects of orthogonal decomposition of the data space for the multi-tiered experiments with respect to different types of multiple randomisation.

The single-phase field experiment described in Section~\ref{sec:Brien1983} is used to demonstrate the orthogonal decomposition method developed by \cite{Brien2009, Brien2010}. The $\mathrm{Treatment}$ term in (\ref{eq:stru3}) is again assumed to be non-orthogonal to both the $\mathrm{Block}$ and $\mathrm{Block.Plot}$ terms in (\ref{eq:stru2expanded}). The two sets of orthogonal projector matrices derived from the structure formulae of the first (block) and second (treatment) tiers are denoted by $\mathcal{Q}\_1$ and $\mathcal{Q}\_2$, respectively. Each orthogonal projector matrix in $\mathcal{Q}\_1$ of the field experiment can be expressed as

\begin{equation}\**label{eq:orthField}**

\{\Q\_{G}, \Q\_{B}, \Q\_{P(B)}\}

\end{equation}

where each orthogonal projector matrix can be computed from the projection matrix defined in Section~\ref{sec:Brien1999}, i.e.\

$\Q\_{G} = \mP\_G$, $\Q\_{B} = \mP\_{B}(\I - \mP\_G) = \mP\_{B} - \mP\_G$ and $\Q\_{P(B)} = (\I - \mP\_{B})(\I - \mP\_G) = \I - \mP\_{B}$. Each of these orthogonal projector matrices in $\mathcal{Q}\_1$ also corresponds to a \emph{stratum} in the ANOVA table of the experiment.

The orthogonal projector matrix in $\mathcal{Q}\_2$ is then

\begin{equation}\**label{eq:orthTrt}**

\{\Q\_{G}, \Q\_{T}\}

\end{equation}

where

$\Q\_{T} = \mP\_{T}(\I - \mP\_G) = \mP\_{T}- \mP\_G$.

The matrix set $\mathcal{Q}\_1$ decomposed by the matrix set $\mathcal{Q}\_2$ is expressed by

\begin{equation}\**label{eq:decomp}**

\mathcal{Q}\_1 \rhd \mathcal{Q}\_2

\end{equation}

which is

\begin{equation}\**label{eq:decompTerm}**

\{\Q\_{G}, \Q\_{B} \rhd \Q\_{T}, \Q\_{B} \vdash \mathcal{Q}\_2, \Q\_{P(B)} \rhd \Q\_{T}, \Q\_{P(B)} \vdash \mathcal{Q}\_2\}

\end{equation}

where $\Q\_{B} \rhd \Q\_{T}$ and $\Q\_{P(B)} \rhd \Q\_{T}$ denote the orthogonal projector matrix for the treatment in the Between and Within Blocks, respectively, while $\Q\_{B} \vdash \mathcal{Q}\_2$ and $\Q\_{P(B)} \vdash \mathcal{Q}\_2$ are the residuals in the Between and Within Blocks. Each element in (\ref{eq:decompTerm}) corresponds to the orthogonal projector matrix of each line in the ANOVA table. The term $\Q\_{B} \rhd \Q\_{T}$, in (\ref{eq:decompTerm}), denotes the Block decomposed by Treatment and then re-expressed as

\[E\_{TB}^{-1}\Q\_{B}\Q\_{T}\Q\_{B} \]

where $E\_{TB}$ denotes the efficiency factor of Block decomposed by Treatment. Another term $\Q\_{B} \vdash \mathcal{Q}\_2$, in (\ref{eq:decompTerm}), denotes the residual of $\mathcal{Q}\_2$ in $\Q\_{B}$ and is then re-expressed as

\begin{equation}\**label{eq:residual}**

\Q\_{B} \vdash \mathcal{Q}\_2 = \Q\_{B} - \sum\_{\Q \in \mathcal{Q}\_2} \Q\_{B} \rhd \Q,

\end{equation}

where $\sum\_{\Q \in \mathcal{Q}\_2}$ denotes the summation over $\Q$ in $\mathcal{Q}\_2$ with non-zero efficiency factor. For the field experiment, the only Treatment term ignores the grand mean in $\mathcal{Q}\_2$; thus, (\ref{eq:residual}) can be written as

\[\Q\_{B} - \Q\_{B} \rhd \Q\_{T}.\]

Hence, each term in (\ref{eq:decompTerm}) can be re-written as

\begin{equation}\**label{eq:decompTermExpaned}**

\{\Q\_{G}, E\_{TB}^{-1}\Q\_{B}\Q\_{T}\Q\_{B}, \Q\_{B} - \Q\_{B} \rhd \Q\_{T}, E\_{T P(B)}^{-1}\Q\_{P(B)}\Q\_{T}\Q\_{P(B)}, \Q\_{P(B)} -\Q\_{P(B)} \rhd \Q\_{T}\}

\end{equation}

where the two efficiency factors $E\_{T B}$ and $E\_{T P(B)}$ should sum to one.

For the multi-tier experiments, the decomposition method is performed across the tiers. Using the wine-evaluation two-phase experiment described in Section~\ref{sec:Brien1983}, three sets of orthogonal projector matrices derived from the structure formulae of each of the three tiers are denoted by $\mathcal{Q}\_1$, $\mathcal{Q}\_2$ and $\mathcal{Q}\_3$, respectively. Each orthogonal projector matrix in $\mathcal{Q}\_1$ of the evaluation experiment can be expressed as

\[\{\Q\_{G}, \Q\_{J}, \Q\_{S(J)}\}, \]

where

$\Q\_{J} = \mP\_{J} - \mP\_G$ and $\Q\_{S(J)} = \I - \mP\_{J}$.

The $\mathcal{Q}\_2$ and $\mathcal{Q}\_3$ consist of the orthogonal projector matrices of the field experiment, which are given previously in (\ref{eq:orthField}) and (\ref{eq:orthTrt}), respectively. The decomposition among the three tiers can be written as

\[(\mathcal{Q}\_1 \rhd \mathcal{Q}\_2) \rhd \mathcal{Q}\_3\]

which equals

\begin{equation}\**label{eq:decompTwoPhase}**

\{\Q\_{G}, \Q\_{J}, \Q\_{S(J)} \rhd \Q\_B, (\Q\_{S(J)} \rhd \Q\_{P(B)}) \rhd \Q\_{T}, (\Q\_{S(J)} \rhd \Q\_{P(B)}) \vdash \mathcal{Q}\_3, \Q\_{S(J)} \rhd \mathcal{Q}\_2\}

\end{equation}

and can be re-written as

\begin{equation}\**label{eq:decompTwoPhaseExpaned}**

\{\Q\_{G}, \Q\_{J}, \Q\_B\Q\_{S(J)}\Q\_B, \Q\_{T} \Q\_{P(B)}\Q\_{S(J)} \Q\_{P(B)}\Q\_{T}, \Q\_{P(B)}\Q\_{S(J)} \Q\_{P(B)} - \Q\_{T}, \Q\_{S(J)} - \Q\_B - \Q\_{P(B)}\}.

\end{equation}

Each of these elements can be inserted into the ANOVA table as shown in Table~\ref{tab:Brien2010}.

\begin{table}[ht]

\centering

\caption{ANOVA table of the wine-evaluation experiment described by \cite{Brien1983}.}

\begin{tabular}{lrll}

\toprule

\multicolumn{1}{l}{\textbf{Source of Variation}} & \multicolumn{1}{l}{\textbf{DF}} & \multicolumn{1}{l}{\textbf{Decomposition}}\\

\midrule

Between Tasters & $t-1$ & $\Q\_{J}$\\

Between Sittings Within Tasters & $t(bp-1)$ &\\

\quad Between Blocks & $b-1$ & $\Q\_B\Q\_{S(J)}\Q\_B$\\

\quad Between Plots Within Blocks & $b(p-1)$ & \\

\quad \quad Treatment & $p-1$ & $\Q\_{T} \Q\_{P(B)}\Q\_{S(J)} \Q\_{P(B)}\Q\_{T}$\\

\quad \quad Residual & $(b-1)(p-1)$ & $ \Q\_{P(B)}\Q\_{S(J)} \Q\_{P(B)} - \Q\_{T}$\\

\quad Residual & $(bp-1)(t-1)$ & $\Q\_{S(J)} - \Q\_B - \Q\_{P(B)}$\\

\bottomrule

\end{tabular}

**\label{tab:Brien2010}**

\end{table}

Different multiple randomisations can have different directions of randomisation between tiers, and \cite{Brien2009, Brien2010} showed that this can affect the ordering of the orthogonal decomposition between terms of different tiers. Additionally, knowing the initial randomisation procedure allows researchers to recognise how terms are confounded with one another between tiers, which can let researchers more easily determine the decomposition method for the experiment. However, categorising types of multiple randomisations for a given multi-tier experiment and then checking how the ordering of the decomposition should be achieved can be laborious. Given that the structural formulae for each tier and for the design are known, a decomposition method presented in Chapter 2 can be used for all two-phase experiments, and does not require categorization of the type of multiple randomisations that should be achieved. Note that the randomisation procedure should still be applied before the experiment to minimise the systematic bias.

Furthermore, \cite{Brien2009, Brien2010} only considered balanced designs, which contain a set of identical efficiency factors for every DF associated with a given block or treatment factor. Chapters 3 and 4 present some optimal designs that are not balanced.

**\section{Recent work by Brien}**

**\label{sec:brien2011}**

More recently, \cite{Brien2011} discussed a systematic approach to designing the two-phase experiments. This paper only considers designs with an orthogonal structure. Some principles described in this study are used to develop the methods used to obtain the optimal designs in Chapters 3 and 4.

\cite{Brien2011} presented a list of rules for calculating the EMS of the ANOVA table. However, these rules can be laborious to follow for experiments that involve numerous treatment and block factors. Chapter 2 of this thesis thus presents an R package that automatically generates the ANOVA table with the EMS for any single- or two-phase experiment.

\cite{Brien2011} then discussed some fundamental principles for designing two-phase experiments. First, the unit of the Phase 1 experiment with the highest variation, but orthogonal to treatments, should be confounded with that of the Phase 2 experiment containing the highest variation. Such a case is illustrated in Chapter 4, which discusses a method of searching for optimal designs where the Phase 1 experiment uses a randomised block design.

Furthermore, if the experimental unit in the Phase 1 experiment has large variation, then consider confounding this unit with one with a smaller variation in the Phase 2 experiment. This approach can minimize the error variance of the estimates of treatment effects as far as possible \citep{Brien2011}.

The treatment can be confounded with multiple units of the Phase 2 experiment, which is caused by the non-orthogonal treatment structure, but this was not discussed by \cite{Brien2011}. Chapters 3 and 4 describe designs that possess the properties of non-orthogonal structures.

**\section{Moving toward high-throughput biotechnologies}**

**\label{sec:Jarrett2008}**

\cite{Jarrett2008} was the first to recognise that experiments of quantitative high-throughput biotechnologies with multiplexing technologies are inherently two-phase experiments. This paper is also the main motivation for this thesis.

\cite{Jarrett2008} conducted a detailed comparative study of the properties of two competing designs, multiple dye-swap and the alternating loop \citep{Churchill2002}, where both designs shared the same size (i.e. same number of replicates of each treatment), for a two-colour microarray experiment (Phase 2) where the Phase 1 experiment had a completely randomised design (CRD). By constructing theoretical analysis of variance (ANOVA) tables for these two designs, they demonstrated that the distribution of the treatment information across the strata of the ANOVA table was dependent on the Phase 2 design. Particularly, while the multiple dye-swap design could be analysed using a simple ANOVA, the alternating loop design required a more involved analysis to test for treatment effects. This is a consequence of the sources of variation introduced in Phase 2 interacting with those introduced in Phase 1. Thus, \cite{Jarrett2008} illustrated the importance of considering the sources of variation introduced at each phase when designing two-phase experiments, and showed that constructing relevant ANOVA tables is a straightforward way to do this.

In some designs, the experimental units from the Phase 1 experiment can be confounded with the array in the Phase 2 experiment. Some of the information from the experimental units can then be estimated in the Between Arrays stratum. However, this information can be recovered, which effectively provides a higher residual DF for error variance of the treatment effects in the Within Arrays stratum. This newly adjusted DF is also known as \emph{effective degrees of freedom} (EDF). \cite{Jarrett2008} demonstrated how to assess the effectiveness of the estimates using the EDF, which is calculated using the first two moments of an approximate chi-square distribution \citep{Satterthwaite1946}. This thesis also uses the EDF to further compare the designs found in Chapters 3 and 4, and the comparison is presented in Chapter 5.

**\section{Proteomic experiment}**

**\label{sec:proteomicExpt}**

This thesis focuses on the design of the MudPIT-iTRAQ$^{\rm TM}$ proteomics experiments. Proteomic studies require the use of a combination of technologies, coupled with database searching, to identify and quantify the proteins within a target cell, tissue or biofluid. This section aims to describe these identifications and measurements in detail, which would help design proteomics experiments using the technologies described. Section~\ref{subsec:protein} thus provides a detailed introduction to proteins and some of their properties, and defines the term proteome. Section~\ref{subsec:MudPIT} describes the process of separating a complex protein mixture into smaller subunits to enable high resolution measurements of the constituent components of proteins. Section~\ref{subsec:iTRAQ} describes a recent protein labelling technology that enables the simultaneous analysis of complex multi-protein mixtures.

**\subsection{Proteins and the Proteome}**

**\label{subsec:protein}**

Proteins are one of the major macromolecules within a biological system that contribute to every process in a living system and are considered essential building blocks of life. They are constructed from chains of amino acids, joined by peptide bonds~\citep{Eidhammer2008}. Amino acid sequences are derived from the genes in the cell nucleus through transcription, where DNA is copied to mRNA, followed by translation, where messenger RNA (mRNA) is decoded into proteins (see Figure~\ref{fig:processes}). Furthermore, a pre-mRNA may undergo alternative splicing as part of the post-transcription process, and thus can produce multiple protein sequences. This makes the study of protein expression more functionally relevant than mRNA transcription, also known as \emph{gene expression}.

A protein can be digested into smaller units, or peptides, that comprise subsequences of the amino acids in the intact protein. Proteins thus are also termed polypeptides, while peptides are produced by enzymatic digestion of a whole protein. An example of such an enzyme is \emph{trypsin}~\citep{Eidhammer2008}.

\begin{figure}[hbt]

\centering{**\includegraphics**[scale=0.55]{image/processes.jpg}}

\caption{Basic biological processes of producing functional proteins from DNA.}

**\label{fig:processes}**

\end{figure}

The proteome is the entire complement of proteins expressed by the genome in a cell, or in tissue or bio-fluid of an organism at a given time under specific conditions \citep{Boehm2007}. The function of a protein corresponds to the timing and location of its expression. Proteomic research thus aims to identify and localise many more unknown and known proteins and better understand their functions.

There are many ways to study proteins. For example, the physical structure of a protein can be studied by X-ray crystallography~\citep{Blow2002}, protein-protein interactions can be studied using the yeast two-hybrid system~\citep{Fields1989}, and the abundance of an individual protein under a defined condition can be studied using isotopic-labelling. The latter is also referred to as \emph{quantitative proteomics}. The use of Multi-dimensional Protein Identification Technology (MudPIT) together with isobaric Tags for Relative and Absolute Quantitation (iTRAQ$^{\rm TM}$) is just one technology used to measure protein abundance.

**\subsection{Multi-dimensional Protein Identification Technology}** **\label{subsec:MudPIT}**

MudPIT is the process of separating a protein complex or peptide mixture using the properties of amino acids, usually into three orthogonal dimensions~\citep{Washburn2001}. Typically, the first separation is by charge, using \emph{strong cation exchange chromatography} (SCX). This is followed by a second separation by hydrophobicity, using \emph{reversed phase liquid chromatography} (RPLC). The third separation is by mass, and is carried out by \emph{mass spectrometry} (MS). This reduces the complexity of the sample and enables high throughput protein analysis. Each MudPIT \emph{run} thus comprises these three steps of separation.

The SCX column contains immobilised negatively charged sulfonic acids that form an ionic interaction with the positively charged peptides. Hence, the charge separation by SCX divides a protein digest (i.e. a mixture of peptides) into many different fractions based on the strength of charge interaction between the peptides and sulfonic acids. Different peptide sequences have different affinities for the SCX resin. This allows for complex peptide mixtures to be fractionated by gradually increasing the concentration of a competing salt solution (for binding to the sulfonic acid groups in a gradient) in a step-wise manner. The salt concentration ranges between 10mM to 500mM. Hence, at each interval of salt concentration, a set of peptides is released into the next MudPIT phase. Each of these charge intervals is termed a \emph{salt step}, and increasing the number of salt steps enables the detection of proteins with low abundance.

The set of peptides from each charge fraction are then separated by RPLC based on hydrophobicity. This is achieved with a separate column that contains silica beads with chains of 18 carbon atoms attached. The peptides are loaded into the column, which causes hydrophobic interactions with the carbon chains. An organic solvent is then added to the column in concentrations that increase over time, causing the peptides to emerge or \emph{elute}, with the least hydrophobic peptides eluting first. Eluted peptides are then detected using the mass spectrometer.

The mass analysis is performed by MS, whereby each peptide's \emph{mass-to-charge ratio} (m/z) is measured and the measurements are then used to calculate the molecular mass. Peptide separation is described in more detail in \cite{Eidhammer2008}.

\emph{Tandem mass spectrometry} (MS/MS), which consists of two repeated phases of MS, was used for this study. Peptide fragmentation occurs between the two phases of MS, and identification and quantification of the peptides and proteins is thus based on these peptide fragments. MS/MS enables higher specificity in protein identification and enables more accurate quantification.

MudPIT has some limitations. For example, large variation in signal intensity between different MudPIT runs can make inter-sample comparisons of peptide or protein abundance difficult. This limitation has been resolved by iTRAQ$^{\rm TM}$ labelling, which enables the simultaneous analysis of up to eight distinct protein digests within a single MudPIT run.

**\subsection{iTRAQ$^{\rm TM}$ for Protein Quantitation}\label{subsec:iTRAQ}**

In its initial format, introduced by \cite{Ross2004}, iTRAQ comprised four isobaric tags, each comprising a reporter group, balance group and peptide reactive group. The reactive group binds the N-terminus at the start of each peptide, and, if the peptide contains lysine residues (i.e. amino acid), then also on the lysine's side chain. The m/z values of the four reporter groups range from 114 to 117, with corresponding balance group values that range from 31 to 28. Each of the four tags thus has an identical total m/z value of 145, making them isobaric. This enables identical peptide species, differentially labelled with the four tags, to be indistinguishable with respect to the intact mass of the peptide when selected for MS/MS~\citep{Ross2004}. For MS/MS, the relative abundances are determined using the reporter ion signals at m/z values of 114, 115, 116 and 117 on the \emph{mass spectrum}. A mass spectrum graphically represents the peptides and peptide fragments based on their m/z values and abundances, and is generated for both phases of MS/MS. The four different labels thus allow the simultaneous analysis of four different samples ~\citep{Ross2004}.

\begin{figure}[htb]

\centering{**\includegraphics**[scale=0.6]{image/iTRAQtags.pdf}}

\caption{Structure of 8-plex-iTRAQ$^{\rm TM}$ tags showing the reporter and balance group masses measured using m/z~\citep{Choe2007}.}

**\label{fig:8-plex}**

\end{figure}

\cite{Choe2007} described a new multiplexing strategy, based on the same concept as the four-plex iTRAQ$^{\rm TM}$ system, that allows the simultaneous analysis of up to eight distinct protein samples (see Figure~\ref{fig:8-plex}). The scheme involves reporter ion signals located at m/z values of 113, 114, 115, 116, 117, 118, 119 and 121. No label is used for an m/z of 120 because it has the same mass as the phenylalanine immonium ion~\citep{Pierce2008}.

**\section{Work by Oberg}**

\cite{Oberg2009} have discussed the experimental design for the mass spectrometry-based proteomic experiments, which include the MudPIT-iTRAQ$^{\rm TM}$ experiments. The goal of an experimental design is to allocate individual samples to iTRAQ tags and MudPIT runs in a way that avoids systematic bias and reduces the error variance of treatment comparison using the ANOVA table.

\cite{Oberg2009} recognised that each MudPIT runs as a block; thus the treatment allocation to runs can be arranged as either a randomised complete block design (RCBD) or a balanced incomplete block design (BIBD). In the RCBD the number of treatments is the same as the number of tags, while in the BIBD the number of treatments exceeds the number of tags. Two additional designs were also discussed: the reference design and loop design. The reference design is where each run contains a reference sample for comparison, while the loop design is where the samples are allocated such that they are cycled through the blocks systematically. \cite{Oberg2009} concluded that the RCBD is most ideal, because it requires fewer runs. Additionally, if a run is lost, this is equivalent to reducing one set of biological replicates. Meanwhile, the resultant design still retains a balanced structure. On the other hand, if the loss of runs occurs in the BIBD and loop design, the resultant design may not preserve the balanced structure. The reference design is most robust under run failure, but may require the largest number of runs for the additional reference sample. Given that the number of tags can be either 4 or 8, then given the high costs of performing the experiment in each MudPIT run, the biologists are likely to utilise all the tags. The RCBD thus may not be possible in all circumstances, because the biologists are unlikely to continuously compare 4 to 8 conditions.

\cite{Oberg2012} again recognised that allocation of treatments to runs and tags is a randomised block design, where the MudPIT run is the block factor. Moreover, \cite{Oberg2012} suggested that multiple MudPIT runs are required to avoid the confounding of tag effects with treatment effects.

Furthermore, the \cite{Oberg2009} and \cite{Oberg2012} only consider the allocation of treatments to runs and tags. The allocation of any other block factors, e.g. animals, is equally important, because it can also affect the error variance of treatment comparison. Therefore, such experiments should always be treated as two-phase experiments.

**\section{Overview}**

**\label{sec:overview}**

The main aim of this thesis is to develop a general construction methodology for two-phase experiments when the Phase 2 experiment involves either a 4- or 8-plex iTRAQ$^{\rm TM}$ proteomics experiment. Such a methodology can be divided into three main parts, as follows.

As mentioned by \cite{Brien2011}, the ANOVA tables with the EMS for two-phase experiments are valuable for comparing the properties of different experimental designs. Chapter 2 first describes a decomposition method for single- and two-phase experiments used to construct the ANOVA table with the EMS. Since no existing tool can automatically generate such tables, an \textsf{R} package called \textbf{infoDecompuTE} is developed that allows the researchers to generate the ANOVA table with EMS by entering any single- or two-phase experimental design.

For a given set of design parameters, there are often many ways to allocate the samples collected from the Phase 1 experiment to the blocks of the Phase 2 experiment. The computational method used to find the optimal designs for different classes of design, such as block, row-column and $\alpha\_n$ designs, has previously been discussed \citep{whitaker1990, Williams1996, John2002}. However, such a method has not yet been developed for optimizing the design of the two-phase experiment. Chapters 3 and 4 thus developed the computer based methods of generating two-phase optimal designs by optimising the objective function using the \emph{simulated annealing algorithm} (SA). The optimality criterion is defined in the \emph{objective function}, which is a mathematical expression generating a numeric value. SA is a well-known heuristic method for finding the input, in this instance the candidate design, that maximizes or minimizes the value generated from this objective function \citep{Kirkpatrick1983}. Additionally, the SA is shown to be further improved in terms of both the speed and efficiency of finding the optimal design. Using the MudPIT-iTRAQ$^{\rm TM}$ experiment as the primary example, the objective function and SA are defined and developed to design an experiment where the Phase 1 experiment involves a completely randomised design (Chapter 3) and a randomised block design (Chapter 4).

Chapter 5 compares several properties of the optimal designs found in Chapters 3 and 4. Firstly, Chapter 5 shows how to estimate the variance components using the restricted maximum likelihood (REML) based on the model that describes the experimental structure of the design. As mentioned in Section~\ref{sec:Jarrett2008}, the information of the experimental units from the Phase 1 experiment can be recovered from the Between Blocks analysis of the Phase 2 experiment, which allows the estimation of the EDF. The EDF, using the variance component estimates, can be computed from the first two moments of the chi-square distribution \citep{Satterthwaite1946}. The comparison of some optimal designs found in Chapters 3 and 4 using the EDF can help clarify the properties of different candidate designs.

Chapter 6 summarises and concludes the entire thesis. Furthermore, it also presents future directions in the design and analysis of the two-phase experiment, particularly for the quantitative high-throughput biotechnologies experiment with multiplexing technology, which is currently useful and will remain so for the next decade.